DOCUMENT RESUME

ED 284 883 TM 870 408

AUTHOR Hathaway, Walter; And Others
TITLE A Regional and Local Item Response Theory Based Test Item Bank System.
PUB DATE Nov 85
NOTE 32p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adaptive Testing; Competency Based Education; Computer Assisted Testing; Criterion Referenced Tests; Elementary Secondary Education; *Item Banks; Language Tests; *Latent Trait Theory; Mathematics Tests; *Minimum Competency Testing; *Quality Control; Reading Tests; *Test Construction; Testing Problems; Test Items; Test Theory
IDENTIFIERS Calibration; Curriculum Alignment; *Fixed Parameter Approach; *Portland School District OR; Rasch Model

ABSTRACT

This report describes the development, operation, maintenance, and future prospects of the item banks pioneered by the Portland (Oregon) School District. At the time of this report, there were 3,500 mathematics, 2,200 reading, and 2,300 language usage items calibrated under the fixed parameter model of item response theory (IRT) for Grades 3-8. The advantages of common scaling of items through IRT methods are discussed, along with particulars for writing new items and conducting controlled field tests. Ongoing concerns include: (1) choosing the best IRT model (one-parameter, used currently, versus two- or three-parameter models); (2) adapting to curriculum dynamics, while preserving item comparability; (3) promoting curriculum alignment of test instruments; (4) making this complex system credible to those outside the testing research field. Twelve quality control guidelines are offered, covering proper printing of tests, test administration procedures, accurate coding and scoring, and retesting criteria. New directions include a pilot project in computerized adaptive testing at elementary and secondary levels and the development of large pools in science, guided by analysis of underlying constructs. The report provides names and addresses of those to contact for access to the item bank system. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED284883.

A REGIONAL AND LOCAL ITEM RESPONSE THEORY BASED TEST ITEM BANK SYSTEM

Walter Hathaway, Ron Houser, and Gage Kingsbury

Portland Public Schools
Department of Research and Evaluation
Portland, Oregon

November 1985

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. Hathaway

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 408

INTRODUCTION

A decade and a half ago, a small group of researchers and test developers in the Portland, Oregon School District realized the potential of Item Response Theory for developing calibrated item banks that would be tools for better educational measurement. The purposes we wanted our citywide testing programs to support included equitable, effective, efficient, valid, and reliable:

- Grouping and placing students
- Targeting instruction on individual student learning needs
- Evaluating student progress over time
- Identifying neglected areas of the aligned curriculum and evaluating and improving programs and services at the student, classroom, grade within school, and grade within district levels
- Providing accountability to the school board and the community

In order to meet these needs, we needed an educational measurement system that would answer the following questions:

- Is the current rate of gain of this student, class, grade, or program satisfactory compared to his/her/its age, grade, program mates and the previous pattern of gains observed?

- What are the current strengths and weaknesses (in terms of goal areas needing further diagnoses and possible work) of this student, class, grade, school, or program, and how have they changed over time?

- Is the level at which this student, class, grade, or program currently performing satisfactory compared to his/her/its age, grade, or program mates and the previous pattern of levels observed?

We could not find any available measurement program that would help us answer these questions and meet these needs adequately, and so we set out to build one ourselves.

There followed a period of extensive collaborative research and development, much of it within the framework of the Northwest Evaluation Association, which was created to foster regional cooperation in and mutual benefit from this effort. The result today is a system of three comprehensive basic skills item banks in Reading, Mathematics, and Language Usage. The constantly growing item banks in Reading and Language Usage each have over 2,000 field tested, calibrated items linked to a common, continuous curriculum scale for each subject. The Mathematics item bank now has over 3,000 such items. State and local school systems have been using these item banks since 1977 to construct effective, efficient survey achievement tests, competency tests, and other instruments that combine the best qualities of criterion referenced and norm referenced measurement. These excellent measurement systems have been the cornerstones of state, district and school renewal efforts that anticipated "A Nation at Risk" (Gardner, 1983) by at least five years. The ongoing collaboration is now resulting in a similar item bank in Science, and yet

another in Social Studies is on the drawing board. Computerized adaptive versions of the tests based upon the item banks are now being pilot tested in Portland.

DESCRIPTION OF THE CURRENT ITEM BANKS

At present, there are approximately 2,200 items that have met all the pre-screening, editing and statistical criteria in Reading; 2,300 items in Language Usage; and 3,500 in Mathematics. Each year 1,800 new items are field tested across the three subject areas, with about 1,200 being added to the item banking systems. Items are added to the banks by a process called the "Developmental Testing Program." Within two weeks of sitting for a regular Achievement Level Test, all students in grades three through eight also sit for a short test of between 20 and 25 items in either Reading, Mathematics, or Language Usage.

In an attempt to more reliably calibrate items on a short test with the current item banks, a procedure is employed that is termed the "fixed parameter model" (Houser, Hathaway & Ingebo, 1983). Since the students have a relatively reliable estimate of their ability through the Achievement Level Testing Program, their ability estimate can be essentially "fixed" for the Development Test at that point on the underlying metric. Item difficulty is the only parameter that is necessary to estimate. Thus, one can calibrate a one item test with the same precision as is achieved on a functional level test.

In cooperation with the Northwest Evaluation Association Science Project, Portland Public Schools is currently in the process of developing a Science item pool which now consists of about 3,500 items.* These items have not been scaled to a common metric at the present time; however, the project is currently engaged in the first phase of a research project that will help determine the number of scales needed and allow the project to begin the design of a series of field tests leading to a final set of scaled Science item banks. The project has plans to scale as many as 16,000 items by the year 1990.

WHY THE ITEM BANKS WERE DEVELOPED

Advantages of Item Response Theory

Classical test theory has been the backbone of educational and psychological testing for most of this century. The concepts of reliability and validity have enabled the development of tests of psychological traits based on sound theory and practical experience. In the last two decades though, testing techniques have been developed and refined to use new theoretical models to describe the interaction between the test-taker and the individual test questions.

---

*The term "pool" was used to distinguish the set of items which do not have a common metric. The term "bank" has been reserved for items which are scaled to a common metric.

These item response theory (IRT: Lord & Novick, 1968; Lord, 1980) models allow the estimation of differences in the measurement characteristics of test questions. These item characteristics may then be used to determine an individual's score on a test and the precision of the score.

The major advantage of these IRT models over classical test theory is that classical item and test characteristics vary, depending upon the group of students taking the test, while IRT item and test characteristics do not. Classical indices of item difficulty, point-biserial correlation, and test-retest reliability may all change when groups of test-takers differ in mean ability or ability distribution. When IRT models are used, item parameters are not biased by the specific sample of students taking the test or by the other items included on the test reference. In addition person parameters are not biased by the particular form of the test taken or by the other students taking the test.

This sample independence means that, except for differences in precision, the student samples used to estimate item characteristics are interchangeable, and the item samples used to estimate student characteristics are interchangeable. It should be noted, however, that these properties of IRT only pertain if the items and students are sampled in some manner from the populations of interest used to derive the IRT scale (Lord & Novick, 1968).

A second advantage of IRT over classical test theory is that a measure of the precision of measurement is available for any question or set of questions with known item characteristics. This standard error differs from that found in classical test theory in that it varies as a function of the student's

achievement level. This is a much more rational result than the constant standard error obtained in classical test theory since we know that the error is not the same for individuals scoring, e.g., 10, 25, and 40 on a 50-item test.

The result of these properties of sample independence and a known functional form for the standard error is that test developers can develop test forms with confidence in the comparability of scores and with some knowledge of the precision of the scores that will be observed. Any set of items taken from a bank of items calibrated using this procedure will produce a scale score and correspondence error of measurement. The size of the error depends on which items are selected, how many items are selected, and the student's raw score on that set of items.

Item Banks and IRT: A Common Scale

A second consequence of the properties of the IRT models is that large groups of items addressing the same subject can be brought together onto a common scale. This makes the IRT models invaluable to anyone trying to create or maintain a large item bank. Large item collections may be easily brought together for use with classical test theory (as in domain-referenced testing); but in order to create anything but randomly parallel test forms, some type of overall scale is a necessity.

The linking procedures and designs which are a direct outgrowth of IRT theory allow researchers to scale thousands of items together (Hathaway, 1980).

From these item banks tests for different purposes and measurement of students' different functional levels may be drawn. Without this common scale the usefulness of a large collection of test questions is severely limited.

THE DEVELOPMENT AND EXPANSION OF THE ITEM BANKS

In the early 1970's, the Portland School District's Research and Evaluation Department, inspired by the work of Georg Rasch and Benjamin Wright, began investigating the applicability of IRT models to the development of item banks from which achievement tests in Reading, Mathematics, and Language Usage could be produced. The Rasch model or one-parameter IRT model was identified as the most promising of these models at the time because of its more advanced stage of development as well as its conceptual clarity and practicality. Later in the decade, while other test developers moved to two- and three-parameter models, Portland's Research and Evaluation Department concentrated its efforts on improving methods of linking items in Rasch-calibrated tests to a single, continuous scale; on reducing the need for second-parameter correction by discarding items that did not fit the model; and on reducing the need for third-parameter correction by readministering higher or lower level tests to students not performing at "percent right" levels that fell within the measuring range of the tests.

Before accepting the Rasch model as a useful tool for the construction of item banks and tests, the Portland Research and Evaluation Department conducted extensive research on such questions as:

1. Does the calibration procedure yield the same scale values regardless of the student sample?

2. How many students are needed to scale test items within acceptable error limits?

3. Does the calibration procedure used produce comparable achievement levels for the same students on different tests?

4. Does an item receive the same scaling regardless of other items in the test?

Once the initial basic research had been completed, work began in earnest on collecting, writing, screening, cataloguing, field testing, calibrating, statistically analyzing, and linking items together into true item banks. Impetus was given to planning the content coverage of the banks by an earlier regional project called the Tri-County Course Development Project which developed, classified, and catalogued comprehensive sets of K-12 learning outcome statements in 14 subject matter areas including those covered by the first three item banks. Collaboration by state and local school districts in Oregon and Washington was gained within the framework of the Northwest Evaluation Association which proved a rich source of ideas, items, field test sites, and other critical, major contributions.

Item Sources

There are two procedures that are employed to identify new items to be field tested. First, items are identified from existing sources such as the National

Assessment of Educational Progress released items, items developed by other federally supported projects, and items shared by other state and local school systems. These sources were an extremely valuable and relatively inexpensive way to get potential new items in the early stages of the item bank development. Prior to the selection of items for each Developmental Testing Program, goal areas which are either lacking depth or some ranges of difficulty are identified. Over time, these "weak" areas have become fewer and thus the specifications for new items have become more restrictive. As this has occurred the public domain item sources have not been as productive a resource as they were in the initial phases of bank development.

Second, items are written to the specifications by trained item writers who usually are teachers. These teachers normally have considerable experience in teaching the level of students for whom they are to write new items. During the item writing process, teachers have the opportunity to critique one another's items. This process helps refine the end product, particularly when writing to a higher level of thinking goals. Considerable care is taken to help the item writers focus on district course goals rather than specific behavioral objectives. It is a policy of the Research and Evaluation Department that specific teaching methodologies not be measured. Rather, the more comprehensive learning goals to which those methodologies are directed are the guiding organization behind the development of the district-wide measurement program. The particular teaching strategies that should be employed in the classroom to help students meet the system's educational goals are left to the professional judgment of the individual teacher.

A number of format specifications are applied to screened and newly written items to ensure good item writing practices and to meet the need to have consistency among the items in a given bank. The sorts of things these guidelines address include: easily understandable directions, clarity, item responses of relatively the same length, good distractors, only one correct answer. If these specifications are not met, the items are either returned to the item writers or item screeners or, if the problem is a minor one, is simply modified by the item bank technical staff.

Items which meet the screening specifications are then assigned an identification number and are sent to the typist. Once the items have been initially formatted on a word processor, they are checked by a diverse team of professionals to determine whether they have any disqualifying gender or ethnic problems identified from an extensive checklist. Problem items are discarded or revised by the team and returned for correction. A final edit is made after all other corrections for any remaining format, spelling, grammar, and/or punctuation errors. The items are then grouped with other items assigned to the same grade level by the item writers or screeners.

Groups of items are then formatted into developmental field tests of 20 to 25 items and administered to students. Once the students' item responses are received, they are matched by student identification numbers to their Achievement Level Test score for that subject. A series of checks are made to determine if the match is correct and if the students did indeed respond to the form of the test indicated on their answer sheets. Upon satisfactory completion of these tests, the "fixed parameter" calibration analysis is run.

Item records are developed for each item on the test containing: the item number on the test, the item bank number, the keyed response, the test form number, the number of omits, the number of students responding to the item, the test date, the Rasch calibration, the calibration standard error, the point biserial correlation, the mean square fit, and the percent of students answering the item correctly. Item characteristic curves are also developed and printed for later evaluation.

The first item screening criterion used following testing is based on the number of students responding to the item. If this number is less than 200, the item is flagged for retesting in the next Developmental Program. The next criteria concerns the percent of students correctly answering the item. If this is less than 5 percent above chance level or greater than 90 percent, the item is flagged for retesting at either a higher or lower grade level as appropriate. In either of these two extremes it is felt that not enough of the item characteristic curve (ICC) is available to reliably determine its calibration. Next, the mean square fit (MSF) is checked, and if there is not a statistically significant misfit, it is then passed to the final statistical check. The last check is a visual inspection of the ICC to identify any abnormal deviations from the ideal model that was not picked up by the MSF.

Any item which fails the first two tests is automatically scheduled for retesting. Any item which fails the last two tests is sent back to a team of item writers for revision if the problem can be identified. If no obvious problem is identified, the item is deleted.

HOW WE CURRENTLY USE THE ITEM BANKS

Test Development

The Reading, Mathematics, and Language Usage item banks are the prime resource employed in the development of the various functional levels tests administered to all students in the Portland Public Schools, grades three through eight, in both Fall and Spring (see Haney, 1985, for a review). Two series of tests in each subject area are designed to measure all six grade levels except in Mathematics, where the instructional sequence of the curriculum dictated the development of two series for each Fall and Spring program so that high achieving younger students would not be expected to respond to items measuring goals for which they have had no instruction.

Each series consists of multiple-level tests of different difficulty. Each level is 20 RITS (2 logits) wide in difficulty. (For students in a particular grade, the standard deviation of achievement levels is about 15 RIT points.) A level's content overlaps the previous and next higher level by 50 percent. A minimum of seven items is necessary to report a subscore and a total test consists of between 45 and 60 items. Students are placed into these levels by either a short locator test, teacher judgment, or by any or all of their previous six scores.

The item banks also comprise the primary resource for the construction of the Graduation Standards Testing Program (Hathaway, 1980). These tests are

designed to make competency decisions and so are focused on a cutoff point on the underlying metric. Most of the items are centered around the cutoff point so that the best measurement is obtained at that point. However, an extension of the measurement downward is achieved by including items which are less difficult. This is done so that goal-based information is available to help identify weak areas for students not performing above the cutoff point.

Computerized Adaptive Testing

To allow more school-based control of testing and to provide school personnel with a capability to use the techniques of modern test theory, we have been investigating the uses of computerized adaptive testing (CAT: Weiss & Kingsbury, 1984).

CAT is a technique for testing which employs a computer to present the most appropriate test questions to a student. The student answers questions on the computer keyboard and the computer scores and records the student's responses. After each item the computer searches an entire item pool (previously stored in the computer) and selects the next test question for the student. At each step in the test, the computer refines its estimate of the student's achievement level and gives the student the question left in the item pool that will provide the maximum information.

From the student's point of view, the test questions become more difficult as questions are answered correctly and less difficult as questions are answered incorrectly. This is akin to the levels assignments done in the functional

levels tests described above, but in a more direct manner, based on the student's current performance.

The CAT system will help serve the schools' special needs for interim measurement, initial pupil placement, and individual graduation standards testing. It is hoped that this form of testing will allow teachers and administrators to assess gains from individualized instruction that occur between the district-wide assessments that occur in the Fall and the Spring. It is expected that the CAT system will serve as a useful addition to the current testing program and help us come one step closer to a continuous measurement model.

The present pilot project, active in six schools ranging from elementary to high schools, is designed to assess the reactions of teachers and students to CAT, and to provide assurance that the CAT system measures the basic skills in the same manner as the current, district-wide, paper-and-pencil tests. Future research will address the potential for enhanced measurement with the CAT system.

PROBLEMS AND ISSUES IN ITEM BANK DEVELOPMENT, MANAGEMENT, AND USE

Choice of an IRT Model

Although the one-parameter model is currently being used with the basic skills item banks, we have not rejected the possibility of the use of other item

response models in the future. As mentioned above, the one-parameter model was originally chosen for use because of its strong theoretical and practical development. If another model were found or developed to the point at which it was demonstrated to have substantial, practical advantage over the one-parameter model, it could be used to recalibrate the current item banks or more likely be applied to the item banks currently under development.

Aside from the one-parameter model based on item difficulty, researchers have developed a two-parameter model which adds an item parameter allowing differences in item discriminatory power and a three-parameter model which adds an item parameter allowing differences in the lowest expected probability of a correct response.

These models differ in three respects. These are the estimation procedures used for item calibration, the scoring procedures used for estimating student achievement levels, and the extent to which student test responses conform to the model.

When building an item bank, it is essential that the response model chosen have practical procedures for the estimation of item and student characteristics. At the same time, the model should be flexible enough to capture the differences inherent in students and items. The one-parameter model has the edge in simplicity of estimation, particularly for student achievement levels. However, the more complex models may result in a more precise representation of item and student performance. Each item bank developer must weigh the advantages of each model with respect to the particular testing situation to determine the best IRT model.

When the basic skills items were first developed, a primary consideration was the number of students that would be needed to calibrate a given set of test questions. It became readily apparent that the only IRT model that would allow stable calibration and linking of items, and not overly restrict the growth of the item pools, would be the one-parameter model. This model allows stable and consistent estimation of item parameters with samples of 200 students. The two- and three-parameter models have required much larger sample sizes to obtain stable calibrations.

Recently, though, Swaminathan and Gifford (1985) have developed new techniques for estimation in the two-parameter model which may reduce the number of students needed to adequately estimate item parameters. In addition Lord (1983) has theoretically demonstrated improved measurement precision using the two-parameter model for all sample sizes greater than 200.

These studies as well as other research findings have led us to begin developing research designed to compare the practical impact of model choice on basic skills measurement.

Dynamism

To insure that the item banks continue to provide measurement reflective of the curriculum as it is being taught at any given time, the banks must be periodically checked and updated to correspond to these changes. There is some concern about the comparability of scores across time with the use of such a dynamic scale. However, metric comparability can be achieved by

employing those subsets of items which have calibrations that remain stable over time. The total measure would still be reflective of the underlying trait at any given point in time.

By adding new items to the item banks after every testing program and by including at least two testing programs in the annual calibration check, the entire item bank would slowly reflect major curriculum shifts. Calibration drifts over the years have generally been modest. The second district-wide administration of the Language Usage test did result in substantial calibration drift, however. To explain this finding, the hypothesis has been advanced that teachers were, for perhaps the first time, focusing instruction on the Language Usage goals.

Interaction with Curriculum

We were fortunate in Portland and in the Northwest in general to have a tradition of a commitment to goal-based instruction and measurement that extended back to the Sixties. We also had a history of strong curriculum and evaluation leadership and collaboration.

One of the greatest challenges, however, in developing the item banks and using them to create good, useful, curriculum aligned, instructionally sensitive measurement instruments has been to gain and maintain effective cooperation between Curriculum and Evaluation. Some of the principles and practices that have helped us succeed in this crucial but delicate part of the work are the following:

. We established and followed the principle that Curriculum would have the leadership and the final say so in determining what was to be measured, while Evaluation would lead on issues of how to measure.

. We shared control and resources. For instance, in Portland we formed a Curriculum and Evaluation Council made up of the senior personnel from each area. This group, which meets monthly, is chaired in alternate years by the heads of the Curriculum and Evaluation Departments. In the meetings information is shared, policy direction is agreed upon, resources are allocated, plans are made and monitored, task forces are created for such work as test blueprint design and reports, revisions, and recommendations of the task forces are reviewed, revised, and ratified.

. We learned how to negotiate (and even enjoy) good, clean conflict since our ends were always the same (namely more informed and better decisions about students, programs, and policies), whereas our disputes were about means which themselves could usually be put to the test of data through experimentation and pilot testing.

. We took the long view. Blessed by enlightened top administrators and school boards, we were able to count on the resources and support to plan and develop on five, ten, and even twenty-year horizons.

Quality Control

Quality control measures are employed to insure that all aspects of the testing program are functioning accurately and effectively. Objectives of quality control and means taken to accomplish them are described below:

1. Insure that schools receive pre-printed answer sheets that are organized as requested by each school (by homeroom, by subject matter, by class, by grade level, etc.) for test administration.

   The school master file showing how each school requested its answer sheets to be organized in the next prior testing is compared with the current set of request forms received from principals. Changes in these requests are identified (usually only a few) and the master file is updated. This file automatically governs the way pre-printed answer sheets are ordered for distribution to the schools.

2. Insure that booklets that are distributed are printed properly, have all pages, etc.

   Samples from each printing order are taken to insure that all pages are present, all items appear in the test, and quality of printing is satisfactory.

3. Insure that the quality of printing on the pre-printed answer sheets will permit accurate scanning.

Pre-printed answer sheets for each test administration are "fan scanned" to find any answer sheets in which the quality of print may not permit accurate scanning, and that all needed information is in its proper place.

4. Insure that the computer program for pre-printing name, I.D., grade level, and test code on answer sheets is functioning accurately.

All programs for pre-printing answer sheets are test-run and results are examined to insure that data are accurate and complete.

5. Insure that students are assigned the proper level of a test series.

The program that assigns students to their proper level tests in any given administration is test-run, and the results for a sample of students are examined in relation to previous test results to insure that the program is functioning properly. (Placement is based on an algorithm averaging each student's five prior test standard scores with double weighting on the last standard score. This average standard score is converted to a RIT score and level placement for the grade level and time of year the testing is being carried out.)

6. Insure that the students understand directions for taking the test.

Verbatim instructions for administering the PALT are provided every teacher, and the testing procedures involve completion of sample items to assure student understanding of how to use the machine-scored answer sheets. There are specially designed answer sheets and practice tests for use with beginning third grade students.

7. Insure that conditions of test taking are good.

Meetings are held with test building coordinators prior to each administration and guidelines for conditions of test taking, as outlined in manuals of instruction for coordinators and school personnel, are addressed at these meetings.

8. Ascertain validity of student performance.

Both the coordinator and teacher manuals state criteria for labelling answer sheets as invalid where student performance does not meet these criteria. In addition computer programs label as invalid any performance falling outside the valid measuring range of the test form the student is assigned to. Also, if a student has 10 or fewer item responses (items attempted) on an answer sheet, the computer notes that performance as invalid.

9. Insure that answer sheets are properly marked and completed.

Each teacher must complete a signed checkoff that he/she has inspected each answer sheet for erasures, multiple marks, extraneous marks, and use of inappropriate markers on the header sheet before the sheets are returned. Scanner programs are used when tests are scored to identify incomplete information for essential hand-coded data.

10. Insure that coded information on each answer is correct and complete.

The procedure described under 3 above insures that pre-coded information is correct.

11. Insure that students have taken the assigned form of the test.

For all students scoring below chance levels, each test is rescored using the scoring keys for all forms in the series. This identifies any student whose test form is different from the one coded on the answer sheet. In such cases the legitimate score is entered on the student's record.

12. Insure that students needing to be retested are retested (high and low).

Students scoring outside the valid measurement range of the test are made known to the school and assigned a higher or lower test form as appropriate (two levels higher or lower, if possible). Schools are provided the materials required for retesting and all schools routinely schedule such retesting following the regular program. Where retests are regarded as inappropriate by teachers, they are not required. In such cases low students receive no test record and high students retain the score achieved.

Complexity and Credibility

Making a complex system credible to the psychometrically unsophisticated is a major issue we have faced in developing IRT-based item banks and functional level testing systems based upon them. This becomes an especially critical

matter when the test results seem to suggest something negative about the performance of students, programs, etc. Rhetoric such as "voodoo testing" is only a breath away when low levels or small gains are registered by complex measurement systems.

There is no magic answer to this problem. Its solution must be approached on a variety of fronts, such as a good inservice program; frequent helpful communication with representatives of the media and with community groups; effective collaboration with public information departments; as well as production and dissemination of a layered system of media productions to tell the measurement system's story--ranging from brochures through workshops and manuals to effective use of electronic media such as video cassettes and cable television. It also helps to periodically have the system evaluated by independent measurement experts who can both sanction the program and suggest improvements.

WHERE WE ARE GOING

New Testing Directions

Since the development of the item bank capability, we have found ourselves continually being asked by teachers, principals, and curriculum specialists for more and better measurement instruments and test results reports. They are developing creative and productive ways to use good data on how students are performing relative to the goals of the district. It is indeed a refreshing change to have the requests for new information come from teachers and principals as well as from the School Board or Superintendent.

For instance, this year 42% of our Black students are reading above the national average up from 36% two years ago and 31% four years ago, an eleven-point gain. Since the national average is the point above which 50% of students achieve, we are now within eight percentage points of having this student population at the national level and are gaining at a rate at which that goal should be reached within three years.

We have been working to develop a scale for measuring attitudes toward the subject of mathematics, and have been exploring ways to create scales for other affective variables. Affective goals are a part of many school districts' curriculum continuums, and yet very little has been done to provide information back to teachers and other instructional leaders as to how effective they have been in developing those traits. We recognize the potential that measurement of affective traits could have in the classroom, but we have only started to develop one such scale and to explore the possibilities for others.

Currently, NWEA and many school districts including Washington, Oregon, California, and Missouri have begun a massive project of developing the capability of measuring Science goals in grades K through 12. Unlike the basic skills item banks, Science has a very diverse curriculum across the participating districts. This diversity produces constraints on the scaling process as well as pointing out a tremendous need to utilize the flexibility of item bank-based measurement by individual school districts.

We are currently examining data gathered in a field testing designed to explore the dimensionality of the underlying traits (Kingsbury, 1985) typically taught in Science. This data will provide the structure to begin

developing the scales for measurement. The five-year measurement objectives of the Science project are to develop as many as 16,000 items distributed across the Science goals. A particularly unique emphasis of the project is to focus not only on content-related items, but also on process and concept goals. Many concepts in Science transcend the particular content in which it is being taught (i.e., conservation, equilibrium cycle, etc.). Developing items to measure these concepts that are independent of the context of a particular subject area is meticulous and thus time consuming.

New Testing Technology

In an effort to allow teachers and school administrators to take advantage of the item banks and modern testing procedures, further applications of computers in the classroom are being explored. As an example, systems are currently being designed to allow the storage and retrieval of test questions from a video disk.

Since the storage capacity of a video disk is so great, this would allow the efficient use of very large item banks. In addition the capability of the video disk to incorporate motion and sound will allow the expansion of the current item banks to include simulation questions, progressive response questions, and questions which do not force the student to read. This will allow the expansion of the CAT program to the more accurate measurement of the basic skills and the expansion of our current measurement to higher level skills.

Another direction of development with the item banks is the technology to allow computerized test design and generation. This new capacity will allow district-wide tests to be developed and produced in a more efficient manner. In addition this technology could be used in the future by teachers and administrators to directly access the item banks to enhance school-based decision making. This use of the computer to store and maintain item content and statistics will make the problem of administering item banks somewhat more manageable.

A basic goal of the item bank work is to close the gap between testing and instruction. One phase of this is to produce tests which have more direct instructional consequence. The CAT pilot project is a first attempt in this direction. By developing tests which more efficiently test the basic skills, more testing time is left to perform pre-diagnostic and diagnostic testing. The ultimate goal of this research effort is to create a more viable interface between the teacher and the evaluator (in this case in the form of the computer).

HOW OTHERS CAN SHARE

Readers who have come this far may be wondering how they can gain access to the item banks, tests, and technology we have been describing and, perhaps, how they can participate in their further development. Here are some answers:

- The current Reading, Mathematics, and Language Usage banks are being disseminated to local school districts for their own use by the Northwest Evaluation Association under a license from the Portland School District. Contact can be made by writing:

Ray Miller, Executive Secretary
NWEA
ESD #121
1410 South 200th
Seattle, WA 98148

or by phoning: 206-242-9400, Ext. 58, or 206-839-3932.

Participation in the ongoing collaboration within the NWEA framework can be initiated through the same source.

- Licenses for use of the current three basic skills item banks by others than local school districts may be requested from the Portland School District by contacting:

Dr. Walter E. Hathaway, Director
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 206

- Information concerning participation in the NWEA Science Curriculum and Assessment project may be obtained by contacting:

Susan Smoyer
NWEA Science Project
700 Pringle Parkway S.E.
Salem, OR 97310
(503) 378-4157

. Questions concerning the Computerized Adaptive Testing project and other applications of modern technology to testing should be addressed to:

Dr. G. Gage Kingsbury
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 229

. Information about the Portland Achievement Levels Tests and the application of IRT to the item banks may be obtained from:

Dr. Ronald L. Houser
Research and Evaluation Department
Portland Public Schools
501 N. Dixon Street
Portland, OR 97227
(503) 249-2000, Ext. 253

. The Portland Achievement Levels Tests and the scoring and reporting software are published and marketed by Microprocessors for Education and Business, 109 N. Main Street, Gresham, OR 97030 (503-666-7883).

. The results of the past and ongoing research and development related to our efforts are available in the form of reports, monographs, conference presentations, articles in popular and refereed journals,

book chapters and books, as well as on electronic networks including the American Educational Research Association's COMPUSERV system and The Council of Great City Schools'. Telenet.

We welcome the interest and involvement of our fellow researchers, developers, and educators in this important work.

References

Gardner, D. P. (Chair) (1983). A Nation at Risk: The Imperative for Educational Reform. Washington, D.C.: U.S. Government Printing Office.

Haney, W. (1985). Making Testing More Educational. Educational Leadership, October, 4-13.

Hathaway, W. E. (1980). A School-District-Developed, Rasch-Based Approach to Minimum Competency Achievement Testing. In R. M. Jaeger & C. K. Tittle (Eds.) Minimum Competency Achievement Testing: Motives, Models, Measures, and Consequences. Beverly Hills, CA: McCutchan.

Hathaway, W. E. (Ed.) (1983). Testing in the Schools: New Directions for Testing and Measurement, Number 19. San Francisco, CA: Jossey-Bass.

Houser, R. L., Hathaway, W. E., & Ingebo, G. S. (1983). An Alternate Procedure to Obtain Ability Estimates in Latent Trait Models. Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

Kingsbury, G. G. (1985). A Comparison of Item Response Theory Procedures for Assessing Response Dimensionality. Paper presented to the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Small N Justifies Rasch Model. In Weiss, D. J. (Ed.) New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. New York: Academic Press.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Swaminathan, H. & Gifford, J. A. (1985). Bayesian estimation in the two-parameter model. Psychometrika, 50, 349-364.

Weiss, D. J. & Kingsbury, G. G. (1984). Applications of Computerized Adaptive Testing to Instructional Problems. Journal of Educational Measurement, 21, 361-375.

0004E
11/21/85